21043

NATIONAL LIBRARY
OTTAWA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR.....CARMEN LEONARD KING.....

TITLE OF THESIS.....A Correlation Study of.....
.....the Intellectual Capabilities of.....
.....the Deaf and Hard of Hearing Students.....

UNIVERSITY.....University of Alberta.....

DEGREE FOR WHICH THESIS WAS PRESENTED.....M.Sc.....

YEAR THIS DEGREE GRANTED.....1974.....

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed).....Carmen King.....

PERMANENT ADDRESS:

.....Clandy Lane,.....
.....C.I. Socorro, San Juan
.....TRINIDAD, West Indies

DATED..9th..Apr.1.....1974

NL-91 (10-68)

THE UNIVERSITY OF ALBERTA

A SIMULATION STUDY OF THE EMPIRICAL DISTRIBUTION OF THE NON-ZERO ROOTS OF SOME SAMPLE COVARIANCE MATRICES

by

© CARMEN LORRAINE KING

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SPRING, 1974

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled A SIMULATION STUDY OF THE EMPIRICAL DISTRIBUTIONS OF THE NON-ZERO ROOTS OF SOME SAMPLE COVARIANCE MATRICES submitted by Carmen King in partial fulfilment of the requirements for the degree of Master of Science.

. . . . . .
Supervisor

. . . . . .

. . . . . .

. . . . . .

Date . . . . . .

ABSTRACT

This thesis is a simulation study of the characteristic roots of the sample covariance matrix. The sample covariance matrices used in the simulation are computed from normally distributed p-variate vectors. The vectors are generated from N(0,1) variates and the population covariance matrix with known eigenvalues. The eigenvalues of the population covariance matrix are arbitrarily chosen. Four non-zero eigenvalues are considered in each case.

The empirical cumulative distribution of each of the non-zero eigenvalues is considered.

The results show that the distribution of the non-zero equal roots show substantial variability. This variability is a function of both the number of equal roots and the rank order of the root. The results for the distinct roots are fairly consistent.

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my Supervisor, Professor Kellogg V. Wilson for his patience, guidance and encouragement throughout the period of my research.

My thanks also the Canadian International Development Agency for their financial support which made this course of study possible.

TABLE OF CONTENTS

	Page
CHAPTER 1 - SOME TESTS BASED ON THE ROOTS OF COVARIANCE MATRICES IN MULTIVARIATE ANALYSIS	1
1.1 Introduction	1
1.2 Tests Involving the Roots of the Single p-variate Matrix	2
CHAPTER 2 - GENERATING AND TESTING THE SIMULATED MODEL	8
2.1 The Maximum Likelihood Estimates of C and μ	8
2.2 Computing a Covariance Matrix with Known Eigenvalues	9
2.3 Computing P the Orthogonal Matrix Used in the Sample	10
2.4 Generating the Normal Rnadom Vectors N(0,C)	11
2.5 Method of Generating the Random Numbers N(0,1)	12
CHAPTER 3 - RESULTS	18
CHAPTER 4 - CONCLUSIONS	44
4.1 Discussion of Results	44
4.2 Comparison of Some Distributions	47
4.3 A Comparison of the Mean, Variance and Skewness of the Same Rank Order and Numerical Value for Four Different Runs	48
BIBLIOGRAPHY	50

LIST OF TABLES

Table		Page
1	Input Data.....	18
2	Mean, Variance and Skewness of the Non-Zero Roots of Run 1.....	21
3	Mean, Variance and Skewness of the Non-Zero Roots of Run 2.....	26
4	Mean, Variance and Skewness of the Non-Zero Roots of Run 3.....	31
5	Mean, Variance and Skewness of the Non-Zero Roots of Run 4.....	35
6	Mean, Variance and Skewness of the Non-Zero Roots of Run 5.....	40
7	Comparison of Some Distributions.....	47
8	Mean, Variance and Skewness of the 4th Root Population Value = 1.0.....	48

LIST OF FIGURES

Figure		Page
1.1	Cumulative Distribution Root 1, Run 1.....	22
1.2	Cumulative Distribution Root 2, Run 1.....	22
1.3	Cumulative Distribution Root 3, Run 1.....	23
1.4	Cumulative Distribution Root 4, Run 1.....	23
1.5	Cumulative Distribution Roots 2 and 3, Run 1.....	24
2.1	Cumulative Distribution Root 1, Run 2.....	27
2.2	Cumulative Distribution Root 2, Run 2.....	27
2.3	Cumulative Distribution Root 3, Run 2.....	28
2.4	Cumulative Distribution Root 4, Run 2.....	28
2.5	Cumulative Distribution Roots 3 and 4, Run 2.....	29
3.1	Cumulative Distribution Root 1, Run 3.....	32
3.2	Cumulative Distribution Root 2, Run 3.....	32
3.3	Cumulative Distribution Root 3, Run 3.....	33
3.4	Cumulative Distribution Root 4, Run 3.....	33
4.1	Cumulative Distribution Root 1, Run 4.....	36
4.2	Cumulative Distribution Root 2, Run 4.....	36
4.3	Cumulative Distribution Root 3, Run 4.....	37
4.4	Cumulative Distribution Root 4, Run 4.....	37
4.5	Cumulative Distribution Roots 2, 3 and 4, Run 4.....	38
5.1	Cumulative Distribution Root 1, Run 5.....	41
5.2	Cumulative Distribution Root 2, Run 5.....	41

Figure		Page
5.3	Cumulative Distribution Root 3, Run 5.....	42
5.4	Cumulative Distribution Root 4, Run 5.....	42
5.5	Cumulative Distribution Roots 2, 3 and 4, Run 5.....	43

GLOSSARY OF SYMBOLS

Symbol	Meaning
$>>$	Much greater than
$\exp(x)$	e^x
$N(0,1)$	Normal with mean 0 and variance 1
A	Matrix A
A'	The transpose of matrix A
A^{-1}	The inverse of matrix A
tr A	The trace of matrix A
$\|A\|$	The determinant of matrix A
$O(f(n))$	Of order not exceeding that of f(n)
$E(x)$	The expected value of X.
$\sum_{i=1}^{n} a_i$	The sum of all elements from a_1 to a_n
$\prod_{i=1}^{n} a_i$	The product of all elements from a_1 to a_n $(a_1 a_2 \ldots a_n)$
I	Identity matrix
P	Number of variables
N	Sample size (number of vectors)

CHAPTER 1

SOME TESTS BASED ON THE ROOTS OF COVARIANCE MATRICES IN MULTIVARIATE ANALYSIS

1.1 Introduction

Multivariate analysis may be defined so the branch of statistical analysis which is concerned with the relationship of sets of variates, that is, the study of vectors of random variables whose components may be correlated with one another.

The main reason for applying multivariate analysis is to solve problems and arrive at numerical results which can be used as the basis for decision making.

The general procedure is as follows:

(i) A hypothesis is proposed.

(ii) Multivariate data is collected.

(iii) A model for the sampling distribution is proposed on the basis of mathematical (usually) results.

(iv) Data is examined and hypothesis tested.

The majority of tests of significance used for testing hypotheses in multivariate analysis are based on the hypothetical sampling distribution of the characteristic roots (eigenvalues) of the multivariate analysis of variance (MANOVA) matrices and the sample covariance matrices.

The tests were derived on the assumption that the random samples are drawn from one or more p-variate normal populations.

Many tests have been formulated for the MANOVA matrices and the joint distribution of the non-zero characteristic roots of these matrices in the null case is given by

$$P(\theta_1 \ldots \theta_k) = C(k,m,n) \prod_{i=1}^{k} \theta_i^m (1-\theta_i)^n \prod_{i>j} (\theta_i - \theta_j)$$

$$(0 < \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k < 1)$$

where $\theta_1, \ldots, \theta_k$ are the non-zero roots of the MANOVA matrices and

$$C(k,m,n) = \frac{\pi^{\frac{1}{2}k} \prod_{i=1}^{k} \Gamma\tfrac{1}{2}(2m+2n+k+i+2)}{\prod_{i=1}^{k} \Gamma\tfrac{1}{2}(2m+i+1)\Gamma\tfrac{1}{2}(2n+i+1)\Gamma\tfrac{1}{2}i}$$

where m,n are interpreted differently for different test situations.

1.2 Tests Involving the Roots of the Single p-variate Matrix

The sample covariance matrix S is given by

$$S = \frac{1}{N-1} \left[\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})' \right] \qquad 1.1$$

where N is the sample size and

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad 1.2$$

S is an unbiased estimate of the population covariance matrix.

$$A = (N-1)S$$

$$= nS$$

The distribution of A (or S) is called the Wishart distribution and is the multivariate generalization of the univariate gamma distribution and therefore plays an important part in statistical inference.

The density function of A for A positive definite is

$$\frac{|A|^{\frac{1}{2}(n-p-1)} \exp(-\frac{1}{2} \operatorname{tr} C^{-1} A)}{2^{np/2} \pi^{p(p-1)/4} |C|^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma\frac{1}{2}(n+1-i)} \qquad 1.3$$

where C is the population covariance matrix.

The joint distribution of the roots of A for C=I is

$$\frac{\pi^{p/2} \prod_{i=1}^{p} \lambda_i^{\frac{1}{2}(n-p-1)} \exp(\frac{1}{2} \sum_{i=1}^{p} \lambda_i) \prod_{i<j} (\lambda_i - \lambda_j)}{2^{pn/2} \prod_{i=1}^{p} (\Gamma[\frac{1}{2}(n+1-i)]\Gamma[\frac{1}{2}(p+1-i)])} \qquad 1.4$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$ are the characteristic roots of A. (Anderson [1]).

Lawley [10] derived tests of significance for the characteristic roots of the sample covariance matrix. He also expressed the expected value $[E(\ell_r)]$ and the variance of the sample roots in terms of the population roots.

Lawley's Tests

Let ℓ_r $(r=1,\ldots,k)$ be one of the k distinct roots of the sample.

(i) If the first k roots of the population covariance matrix are distinct, and the remaining p-k roots are equal to δ say, to test the hypothesis of equality of the p-k roots, the test statistic is given by

$$\text{Const}\{-\log_e(\ell_{k+1}\ldots\ell_p)/\delta^{p-k} + (\ell_{k+1}+\ldots+\ell_p)/\delta - (p-k)) \quad 1.5$$

which is approximate χ^2 with $\frac{1}{2}(p-k)(p-k+1)$ degrees of freedom.

$$\text{Const} = n - k - \frac{1}{6}(2q+1-\frac{2}{q+1}) - \frac{1}{q+1}\{\sum_{r=1}^{k}\frac{\lambda_r}{\lambda_r-\delta}\} +$$

$$\delta^2 \sum_{r=1}^{k}\frac{1}{(\lambda_r-\delta)^2} \quad 1.6$$

If $\lambda_1\ldots\lambda_k >> \delta$ Const is given by

$$n - k - \frac{1}{6}(2q+1-\frac{2}{q+1}) - \frac{k^2}{q+1}$$

where $q = p-k$.

(ii) If δ is unknown the approximate χ^2 has $\frac{1}{2}(p-k-1)(p-k+2)$ degrees of freedom and the test of the hypothesis becomes

$$\text{Const}\{-\log_e(\ell_{k+1}\ldots\ell_p) + (p-k)\log_e(\ell_{k+1}+\ldots+\ell_p)/(p-k)\} \quad 1.7$$

where

$$\text{Const} = n - k - \frac{1}{6}(2q+1+\frac{2}{q}) + \delta^2 \sum_{r=1}^{k} \frac{1}{(\lambda_r-\delta)^2} \quad 1.8$$

If $\lambda_r >> \delta$

$$\text{Const} = n - k - \frac{1}{6}(2q+1+\frac{2}{q})$$

When k=0 this reduces to the hypothesis of equality of the roots and Const is then given by

$$n - \frac{1}{6}(2p+1+\frac{2}{p}) \quad 1.9$$

(iii) If ℓ_r is one of the k distinct roots $r=1,\ldots,k$

$$E(\ell_r) = \lambda_r + \frac{\lambda_r}{n} \sum_{\substack{i=1 \\ \neq r}}^{p} \left(\frac{\lambda_i}{\lambda_r-\lambda_i}\right) + O\frac{1}{n^2} \quad 1.10$$

and the variance of ℓ_r is

$$\frac{2\lambda_r^2}{n}\{1 - \frac{1}{n}\sum_{\substack{i=1 \\ \neq r}}^{p}(\frac{\lambda_i}{\lambda_r-\lambda_i})^2\} + O\frac{1}{n^3} \qquad 1.11$$

Where O(f(n)) signifies an expression order not exceeding f(n).

In 1.10 and 1.11, n is the sample size (number of vectors), if n is large $O(n^{-2})$ and $O(n^{-3})$ are insignificant.

The above expressions are valid only if the roots are distinct, there are no available expressions for the non-zero equal roots.

It follows from 1.10 and 1.11 that the mean and variance and hence the distribution of each of the non-zero distinct roots depend on the other characteristic roots, but the extent of the dependence cannot be exactly determined from 1.10 and 1.11 as these expressions are not exact.

To construct exact significance tests, and to be aware of the magnitude of possible sampling errors it is important to know the sampling distributions of the estimates of these roots.

In this simulation study the sample chosen for the study of the empirical distributions of the sample roots is the single sample p-variate covariance matrix.

The following points would be considered in this thesis:

(1) The empirical cumulative distribution of each of the non-zero characteristic roots (population roots known).

(i) When the non-zero roots are distinct.

(ii) When the non-zero roots contain equal roots.

(2) A comparison of the distributions of some of the roots from different computer runs.

(3) The general behaviour of the non-zero equal roots and the dependence of their distribution on the rank order of the roots.

CHAPTER 2

GENERATING AND TESTING THE SIMULATED MODEL

2.1 The Maximum Likelihood Estimates of C and μ

The model generated is based on the theory that the p-variate vectors used to compute the sample covariance matrix S are distributed N(0,C) where C, the population covariance matrix is known.

The density function of the multivariate normal is

$$\frac{1}{2\pi^{p/2}\ |C|^{\frac{1}{2}}} \exp[\tfrac{1}{2}(X-\mu)'\ C^{-1}(X-\mu)] \qquad 2.1$$

where

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

and μ is a vector of means.

The maximum likelihood estimates of C and μ are given by

$$A = \frac{1}{N}\left[\sum_{i=1}^{N} (x_i-\bar{x})(x_i-\bar{x})'\right]$$

and from 1.2

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

respectively.

$S = \frac{N}{N-1} A$ is the unbiased estimate of C.

2.2 Computing a Covariance Matrix with Known Eigenvalues

For every real symmetric matrix B, there exists an orthogonal matrix P such that

$$P'BP = D \qquad 2.2$$

where D is the diagonal matrix of the eigenvalues of B.

If C is the covariance matrix (real and symmetric) with eigenvalues

$$\lambda_1, \lambda_2, \ldots, \lambda_p$$

then

$$P'CP = \Lambda$$

where Λ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$. Premultiplying by P and post-multiplying by P' (since P is orthogonal)

$$C = P\Lambda P'$$

$$= A_1 A_1' \quad (C \text{ is a symmetric matrix})$$

$$= P(\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}})P'$$

$$= (P\Lambda^{\frac{1}{2}})(\Lambda^{\frac{1}{2}}P')$$

$$= (P\Lambda^{\frac{1}{2}})(P\Lambda^{\frac{1}{2}})' \qquad 2.3$$

$$\therefore A_1 = P\Lambda^{\frac{1}{2}} \qquad 2.4$$

and $\Lambda^{\frac{1}{2}}$ is the diagonal matrix of the square root of the eigenvalues of C.

2.3 Computing P the Orthogonal Matrix Used in the Sample

An arbitrary matrix A_0 was chosen, such that

$$A_0 = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.7 & 0.7 & 0.1 & 0.1 \\ 0.6 & 0.8 & 0.0 & 0.0 \\ 0.3 & 0.3 & 0.9 & 0.1 \\ 0.5 & 0.5 & 0.1 & 0.7 \\ 0.0 & 0.0 & 0.8 & 0.6 \\ 0.1 & 0.1 & 0.7 & 0.7 \\ 0.8 & 0.4 & 0.4 & 0.2 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.9 & 0.3 & 0.3 & 0.1 \end{bmatrix}$$

so that $A_0 A_0'$ is a symmetric matrix.

The eigenvalues and eigenvectors of $A_0 A_0'$ were then

calculated. The eigenvectors calculated were all orthogonal and these p eigenvectors formed the orthogonal matrix P. Having found P, the covariance matrix C was then computed using 2.3 and 2.4

In the first case the eigenvalues of C, λ_i were chosen to be 5,2,2,1,...,0 for i=1,p and

$$\Lambda^{\frac{1}{2}} = \begin{bmatrix} \lambda_1^{\frac{1}{2}} & & & & & 0 \\ & \lambda_2^{\frac{1}{2}} & & & & \\ & & \lambda_3^{\frac{1}{2}} & & & \\ & & & \lambda_4^{\frac{1}{2}} & & \\ & & & & \ddots & \\ 0 & & & & & \lambda_p^{\frac{1}{2}} \end{bmatrix}$$

$$C = (P\Lambda^{\frac{1}{2}})(P\Lambda^{\frac{1}{2}})'$$

2.4 Generating the Normal Random Vectors N(0,C)

From 2.3 we have

$$C = A_1 A_1'$$

Let Y be distributed N(0,I) where I_{pxp} is the identity matrix, and let

$$X = A_1 Y$$

Then X is distributed $N(0, A_1 A_1')$

$$x_i = \sum_{j=1}^{p} a_{1ij} y_j \quad (i=1,\ldots,p) \qquad 2.5$$

where a_{1ij} is the ij^{th} element of A_1 and $y_1,\ldots,y_p$ are p independent standard normal variables N(0,1).

2.5 Method of Generating the Random Numbers N(0,1)

The algorithm formulated by Chen [6] was used to generate the N(0,1) numbers.

This algorithm generates pseudo-random numbers for a 32-bit word computer, for example the IBM 360/67 which was used in the simulation.

The theory is based on the multiplicative congruential method given by

$$R_i = R_{i-1}(2^p+k) \pmod{2^{31}} \qquad 2.6$$

where p is a positive integer greater than 2 and less than 31, and k is any odd integer.

The random deviates of the unit uniform distribution

are obtained by putting

$$U_i = R_i/(2^{31}-1) \qquad 2.7$$

finally, two independent random normal deviates of mean 0 and variance 1 are produced from two independent uniform deviates U_1 and U_2 by the transformation suggested by Box and Muller [5] that is

$$Y_1 = (-2 \log_e U_1)^{\frac{1}{2}} \cos 2\pi U_2 \qquad 2.8$$

$$Y_2 = (-2 \log_e U_1)^{\frac{1}{2}} \sin 2\pi U_2 \qquad 2.9$$

Chen chose to generate the two random uniform deviates at a time rather than one. Using the multiplicative congruential method this becomes

$$R_{1i} = R_{1,i-1}(2^{p_1}+k)(\text{mod } 2^{31}) \qquad 2.10$$

$$R_{2i} = R_{2,i-1}(2^{p_2}+k)(\text{mod } 2^{31}) \qquad 2.11$$

Each generation of this dual type produces alternate numbers for the sequence.

It was found that generators with k=3 perform better

than those with k=1. (2^p+k) is relatively prime to 2^{31} and is an odd number, therefore k is an odd number. Chen used a value of k=3 in his generator.

Values of p_1=14 and p_2=18 were found empirically by Chen to be the best among all possible combinations of $6 \leq p_1, p_2 \leq 18$. He further improved the generator by changing 2.11 to give

$$R_{2i} = [R_{2,i-1}(2^{18}+3)](2^{18}+3)(\text{mod } 2^{31}) \qquad 2.12$$

He used a sample size of 10^6 random numbers to test the generator and the results of the tests performed seem satisfactory for the purpose of this simulation.

10,000 random N(0,1) numbers were generated using Chen's algorithm for each of five different sets of starting values. The Kolmogorov-Smirnov test statistic was then used to test the goodness of fit of the sample to the N(0,1) distribution. Each sample set was divided into 162 intervals from -4 to 4. The maximum deviation from among all the sets was 0.008, which is in agreement with the results obtained by Chen.

At the 5 percent level of significance the critical value of the Kolmogorov-Smirnov test is 0.0136.

The hypothesis that the numbers generated are N(0,1) was therefore accepted at the five percent level of significance.

The Model

A matrix Y(pxN) of N(0,1) variates was generated with p=10 and N=100

$$Y = \begin{bmatrix} y_{11} & y_{1N} \\ y_{21} & y_{2n} \\ \vdots & \vdots \\ y_{p1} & y_{pN} \end{bmatrix} \qquad 2.18$$

and using A_1 from 2.4, the matrix of random vectors was computed from

$$A = A_1 Y$$

and putting $X_i = (X_i - \bar{x})$

where $$\bar{x} = \frac{1}{N} \begin{bmatrix} \sum_{i=1}^{N} & x_{1i} \\ \vdots & \\ \sum_{i=1}^{N} & x_{pi} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

S was then compuated as

$$\frac{1}{N-1}\left[\sum_{i=1}^{N}(X_i)(X_i)'\right]$$

Testing the Goodness of Fit of the Model

1,000 samples of S were generated from a population covariance matrix with four non-zero characteristic roots, and for each run the test statistic given by

$$n - \frac{1}{6}(2P+1 - \frac{2}{p+1})(\log\frac{|C|}{|S|} - p + \text{tr } C^{-1}S) \qquad 2.15$$

(which is approximately χ^2 with $\frac{1}{2}(p)(p+1)$ degrees of freedom was calculated.

In order to compute this statistic some assumptions had to be made, since the zero roots in the population covariance matrix C would render the statistic indeterminate. In all the goodness of fit tests that have been derived so far, the test statistic is dependent on the product of the population roots or estimates of them.

In factor analytic methods, after the appropriate number of factors have been fitted, the remaining factors tend to be zero. This being the case 10^{-6} was considered a reasonable approximation for the zero roots.

The 1,000 values of the test statistic were divided into 54 intervals from 28.0 to 97.0 and the Kolmogorov-Smirnov test was used to test the goodness of fit with a χ^2

distribution with 55 degrees of freedom. The maximum difference was 0.031, at the five percent level of significance the critical value of the test is 0.043. Since the maximum deviation is below this value, the hypothesis that the sample covariance matrix S is a good representation of the population covariance matrix C was accepted at the five percent level of significance.

CHAPTER 3

RESULTS

Five Computer runs were made each with a population covariance matrix having four non-zero roots. The non-zero characteristic roots of the population matrix were arbitrarily chosen as follows:

TABLE 1

INPUT DATA

Rank Order of Root	Value of the Population Roots				
	Run 1	Run 2	Run 3	Run 4	Run 5
1	5	5	4	7	4
2	2	3	3	1	2
3	2	1	2	1	2
4	1	1	1	1	2

Since various test procedures require that the characteristic roots be in descending order of magnitude, the roots were arranged in this order.

1,000 samples were generated for each run with n=00, p=10, N_1=1000.

The sample mean, variance and skewness were calculated for each root using

$$\text{Sample mean } \bar{\ell} = \frac{1}{N_1} \sum_{i=1}^{N_1} \ell_i$$

where ℓ_i is a root of the sample

$$\text{Sample variance } S^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (\ell_i - \bar{\ell})^2$$

and

$$\text{Skewness} = \sum_{i=1}^{N_1} \frac{(\ell_i - \bar{\ell})^3}{(N_1 - 1)S^3}$$

$E(\ell_r)$ and $\text{Var}(\ell_r)$ were also calculated using 1.10 and 1.11. Since the sample used sample size N=100, $\frac{1}{n^2}$ and $\frac{1}{n^3}$ are of the order 10^{-4} and 10^{-6} respectively, these values are insignificant compared with the other values. These terms were therefore omitted in the calculations.

Computer Run 1

The Population Covariance Matrix

$$C = \begin{bmatrix}
1.159 & & & & & & & & & \\
0.535 & 0.844 & & & & & & & & \\
0.387 & 0.842 & 0.888 & & & & & & & \\
0.192 & 0.424 & 0.361 & 1.094 & & & & & & \\
0.369 & 0.654 & 0.580 & 0.119 & 1.334 & & & & & \\
-0.046 & 0.100 & -0.018 & 0.664 & 0.436 & 0.880 & & & & \\
0.034 & 0.219 & 0.099 & 0.571 & 0.651 & 0.863 & 0.902 & & & \\
0.752 & 0.731 & 0.632 & 0.643 & 0.585 & 0.353 & 0.421 & 0.860 & & \\
-0.385 & 0.651 & 0.820 & 0.307 & 0.449 & 0.012 & 0.098 & 0.226 & 1.314 & \\
0.913 & 0.705 & 0.591 & 0.556 & 0.495 & 0.224 & 0.288 & 0.866 & 0.055 & 0.925
\end{bmatrix}$$

Since the matrix is symmetric only the lower triangular portion is given. The values are rounded to three decimal places for each computer run.

TABLE 2

MEAN, VARIANCE AND SKEWNESS OF THE NON-ZERO ROOTS OF RUN 1

Pop. Value of Root	Sample Mean	Sample Variance	$E(\ell_r)$	$Var(\ell_r)$	Skewness
5.0	5.0805	0.5063	5.0800	0.495	0.1717
2.0	2.2167	0.0676	--	--	0.4657
2.0	1.7256	0.0454	--	--	0.0914
1.0	0.9466	0.0176	0.9475	0.0179	0.1711

The expressions for $E(\ell_r)$ and $Var(\ell_r)$ are not applicable when the non-zero roots are equal.

Figs. 1.1, 1.2, 1.3, 1.4 show the cumulative distribution of each of the non-zero roots.

Fig. 1.5 shows the distribution of roots 2 and 3 when the population root = 2.0.

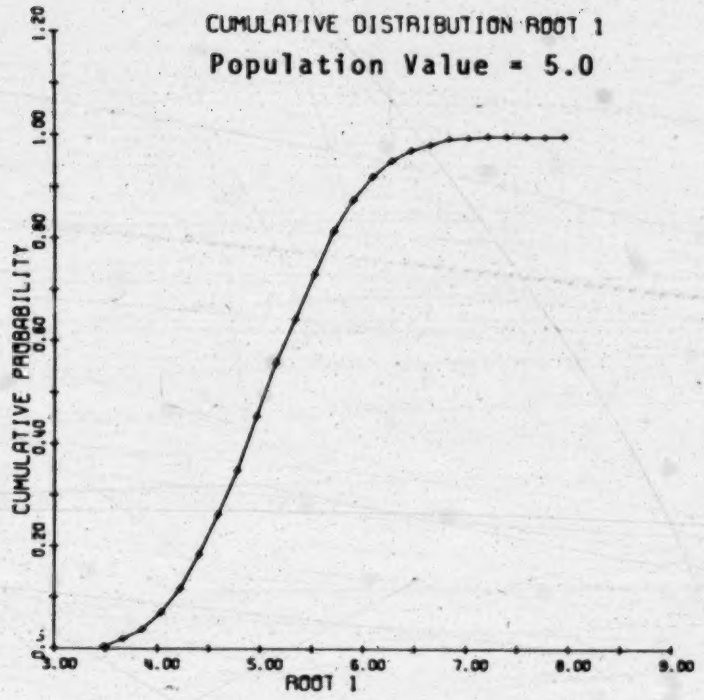
CUMULATIVE DISTRIBUTION ROOT 1
Population Value = 5.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
4.00
5.00
6.00
7.00
8.00
9.00
ROOT 1

Fig. 1.1

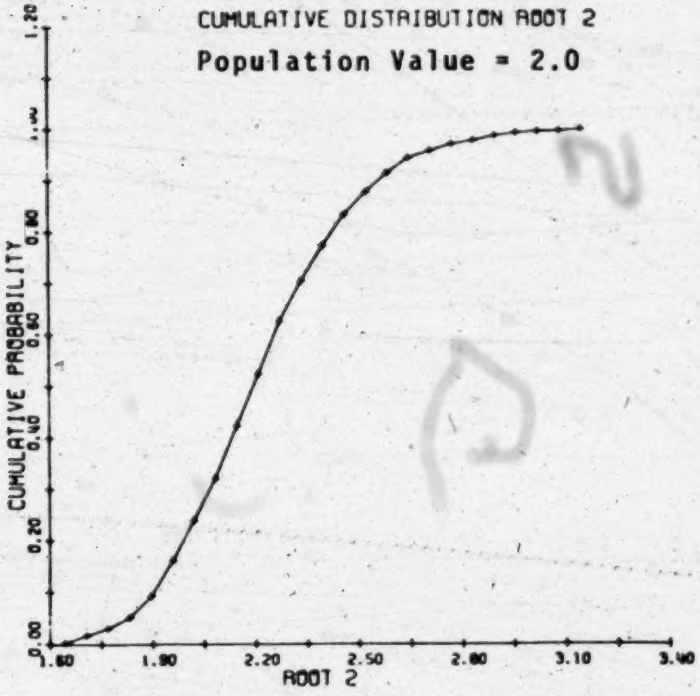
CUMULATIVE DISTRIBUTION ROOT 2
Population Value = 2.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.20
1.80
1.90
2.20
2.50
2.80
3.10
3.40
ROOT 2

Fig. 1.2

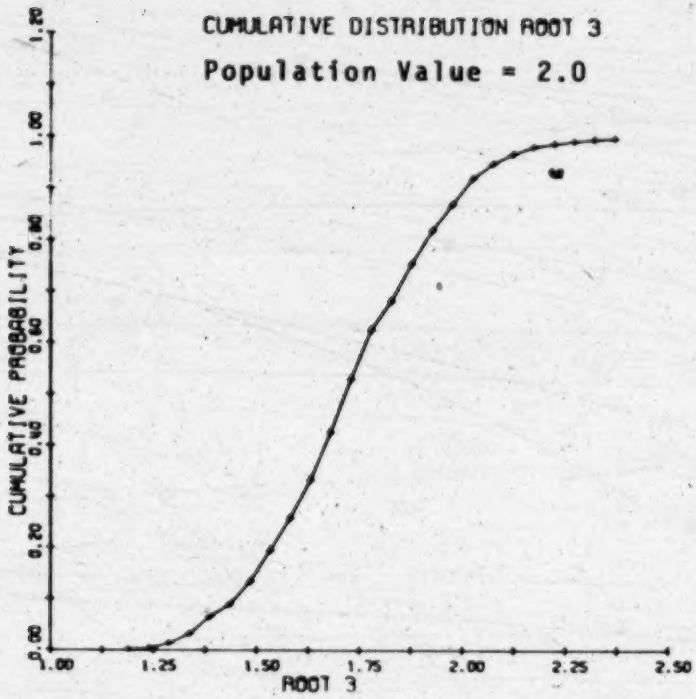
CUMULATIVE DISTRIBUTION ROOT 3
Population Value = 2.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.00
1.25
1.50
1.75
2.00
2.25
2.50
ROOT 3

Fig. 1.3

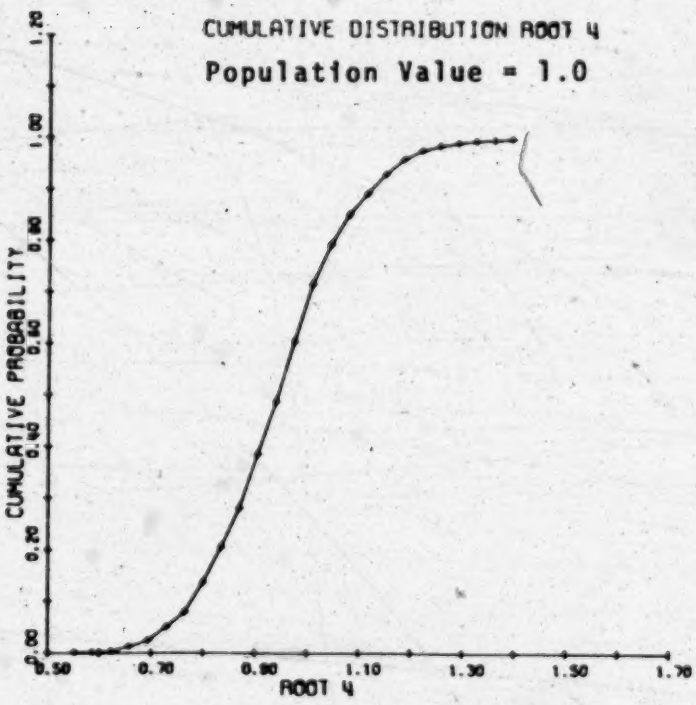
CUMULATIVE DISTRIBUTION ROOT 4
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
0.50
0.70
0.80
1.10
1.30
1.50
1.70
ROOT 4

Fig. 1.4

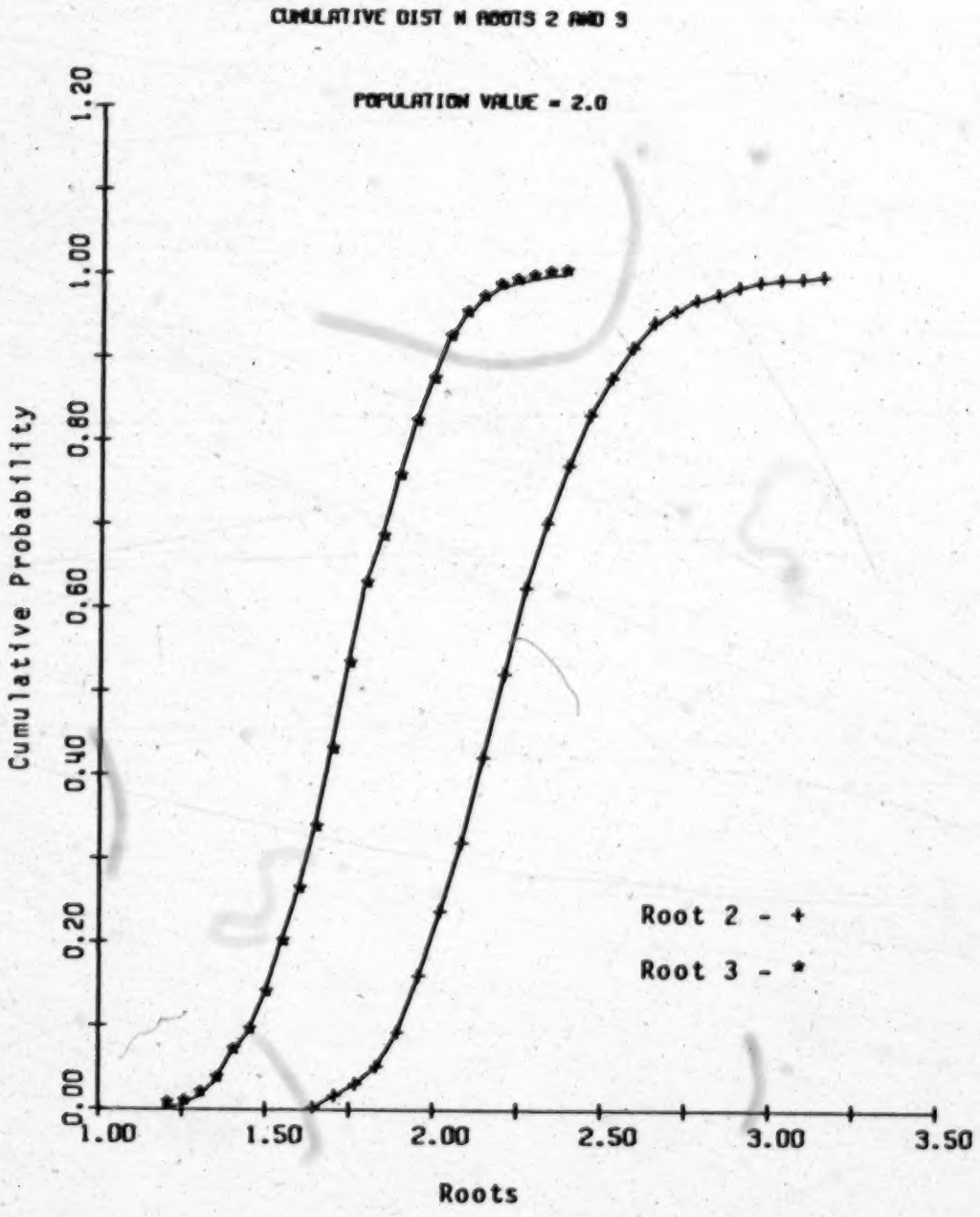
CUMULATIVE DIST N ROOTS 2 AND 3
POPULATION VALUE = 2.0
Cumulative Probability
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.00
1.50
2.00
2.50
3.00
3.50
Roots
Root 2 - +
Root 3 - *

Fig. 1.5

Computer Run 2

The Population Covariance Matrix

$$C = \begin{bmatrix}
0.885 & & & & & & & & & \\
0.643 & 0.893 & & & & & & & & \\
0.581 & 0.899 & 0.947 & & & & & & & \\
0.137 & 0.342 & 0.248 & 1.199 & & & & & & \\
0.428 & 0.629 & 0.537 & 0.132 & 1.121 & & & & & \\
-0.152 & -0.044 & -0.216 & 0.853 & 0.461 & 1.219 & & & & \\
-0.046 & 0.087 & -0.084 & 0.740 & 0.670 & 1.165 & 1.171 & & & \\
0.661 & 0.750 & 0.672 & 0.643 & 0.605 & 0.351 & 0.424 & 0.833 & & \\
0.062 & 0.641 & 0.747 & 0.207 & 0.350 & -0.155 & -0.071 & 0.344 & 0.888 & \\
0.764 & 0.756 & 0.686 & 0.535 & 0.528 & 0.181 & 0.258 & 0.818 & 0.285 & 0.844
\end{bmatrix}$$

TABLE 3

MEAN, VARIANCE AND SKEWNESS OF THE NON-ZERO ROOTS OF RUN 2

Pop.	Root	Sample Mean	Sample Variance	$E(\ell_r)$	$Var(\ell_r)$	Skewness
5		5.1013	0.5042	5.1010	0.488	0.2017
3		2.9170	0.1641	2.954	0.169	0.3287
1		1.0929	0.0154	--	--	0.4710
1		0.8483	0.0122	--	--	0.1441

Figs. 2.1, 2.2, 2.3 and 2.4 show the cumulative distribution of each of these roots.

Fig. 2.5 shows the cumulative distribution of roots 3 and 4.

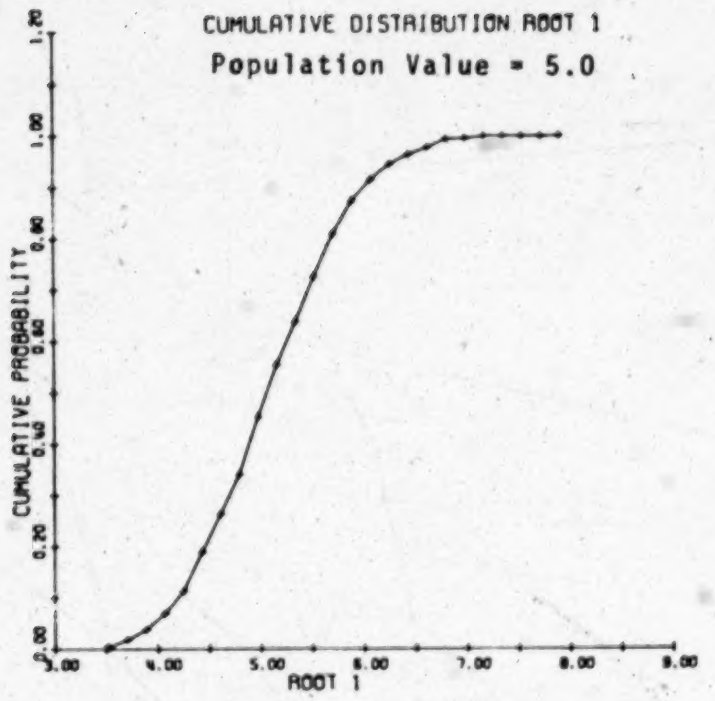
CUMULATIVE DISTRIBUTION ROOT 1
Population Value = 5.0
CUMULATIVE PROBABILITY
ROOT 1

Fig. 2.1

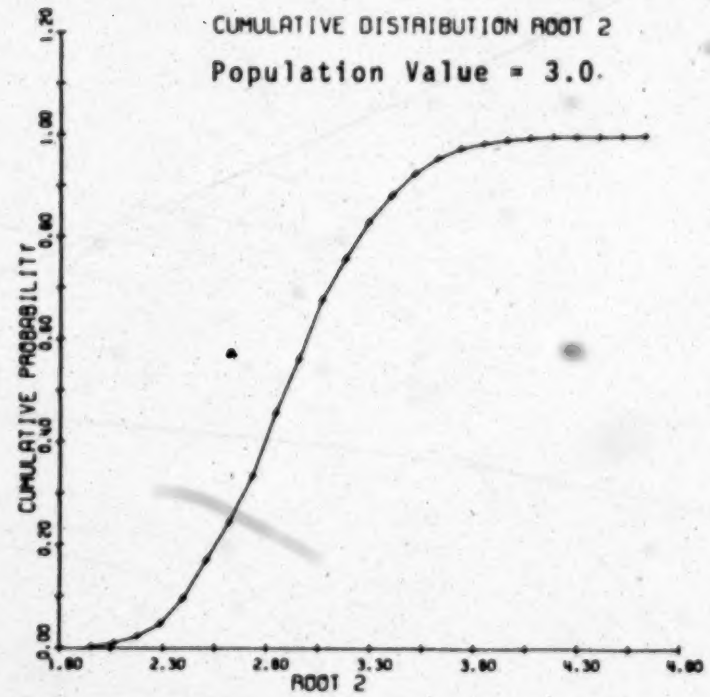
CUMULATIVE DISTRIBUTION ROOT 2
Population Value = 3.0
CUMULATIVE PROBABILITY
ROOT 2

Fig. 2.2

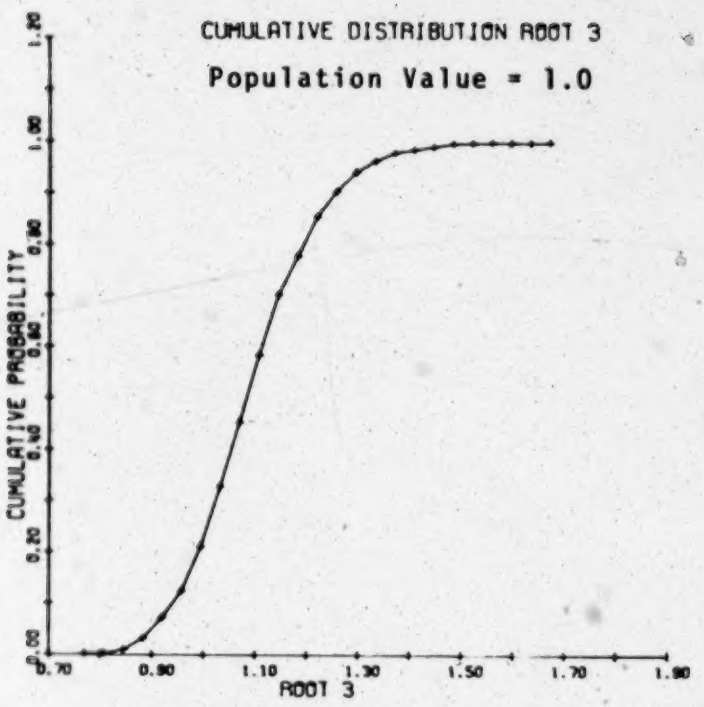
CUMULATIVE DISTRIBUTION ROOT 3
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.70
0.90
1.10
1.30
1.50
1.70
1.90
ROOT 3

Fig. 2.3

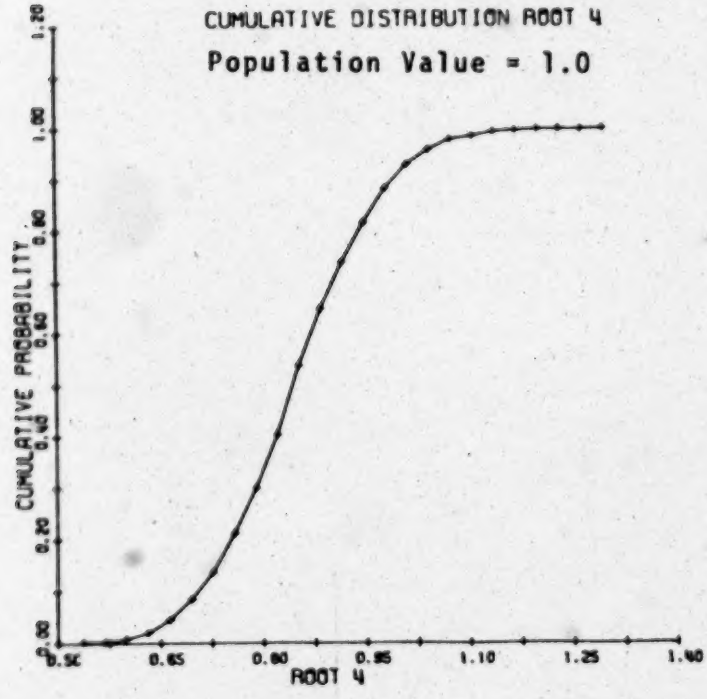
CUMULATIVE DISTRIBUTION ROOT 4
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.50
0.65
0.80
0.95
1.10
1.25
1.40
ROOT 4

Fig. 2.4

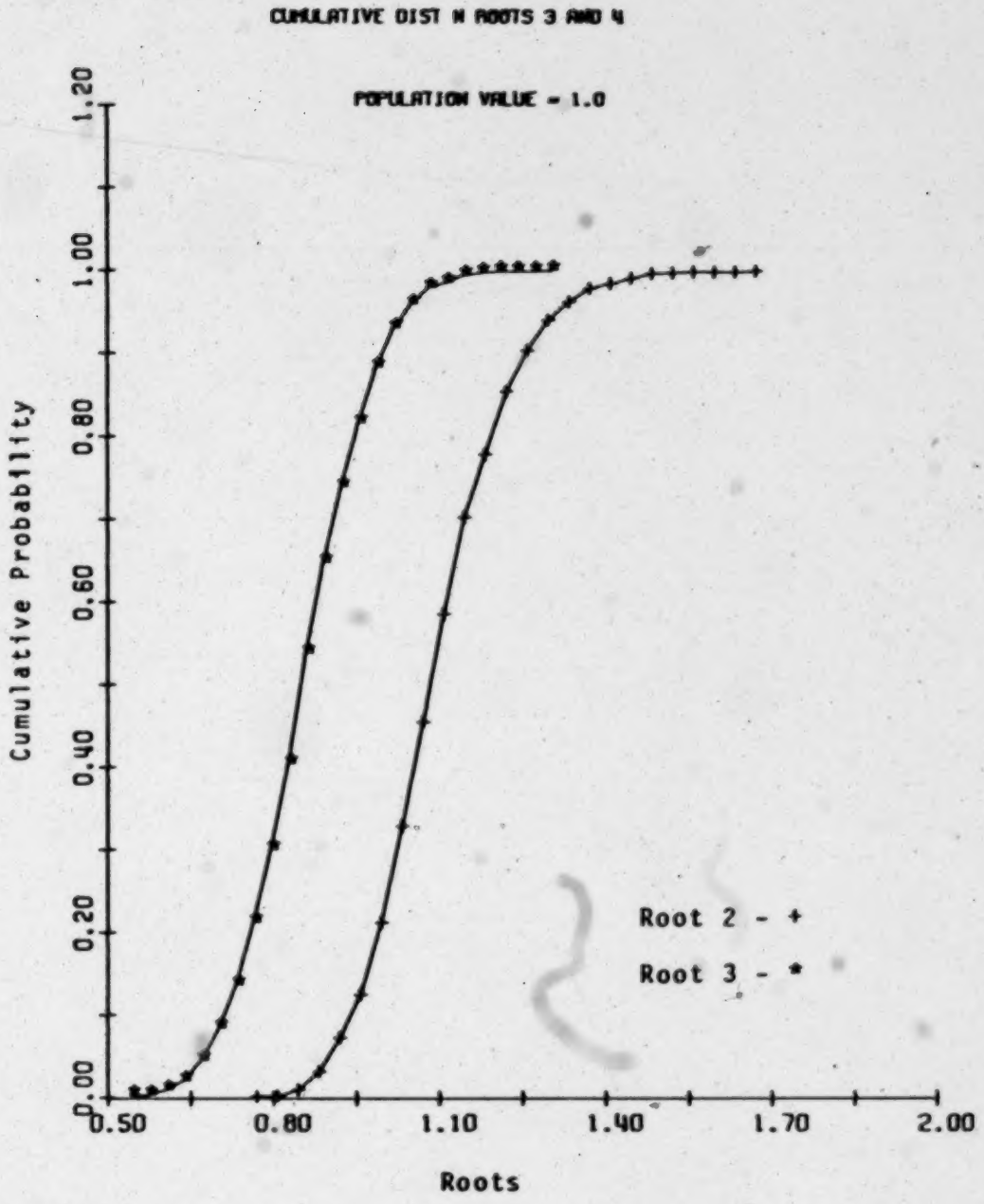
CUMULATIVE DIST N ROOTS 3 AND 4
POPULATION VALUE = 1.0
Cumulative Probability
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.50
0.80
1.10
1.40
1.70
2.00
Roots
Root 2 - +
Root 3 - *

Fig. 2.5

Computer Run 3

The Population Covariance Matrix

$$C = \begin{bmatrix}
1.132 & & & & & & & & & \\
0.484 & 0.759 & & & & & & & & \\
0.364 & 0.789 & 0.873 & & & & & & & \\
0.031 & 0.231 & 0.152 & 1.109 & & & & & & \\
0.258 & 0.509 & 0.444 & 0.031 & 1.015 & & & & & \\
-0.235 & -0.114 & -0.273 & 0.796 & 0.398 & 1.182 & & & & \\
-0.155 & 0.008 & -0.153 & 0.669 & 0.594 & 1.121 & 1.117 & & & \\
0.644 & 0.584 & 0.500 & 0.516 & 0.445 & 0.265 & 0.316 & 0.701 & & \\
-0.394 & 0.623 & 0.818 & 0.166 & 0.361 & -0.165 & -0.742 & 0.142 & 1.318 & \\
0.829 & 0.586 & 0.493 & 0.410 & 0.359 & 0.092 & 0.145 & 0.717 & -0.005 & 0.792
\end{bmatrix}$$

TABLE 4

MEAN, VARIANCE AND SKEWNESS OF THE NON-ZERO ROOTS OF RUN 3

Pop. Root	Sample Mean	Sample Variance	$E(\ell_r)$	$Var(\ell_r)$	Skewness
4	4.1834	0.3011	4.1740	0.2904	0.2756
3	2.9189	0.1336	2.9545	0.1446	0.2728
2	1.9029	0.0654	1.9192	0.0693	0.1939
1	0.9512	0.0180	0.9512	0.0186	0.2053

Figs. 3.1, 3.2, 3.3 and 3.4 show the cumulative distribution of each of these roots.

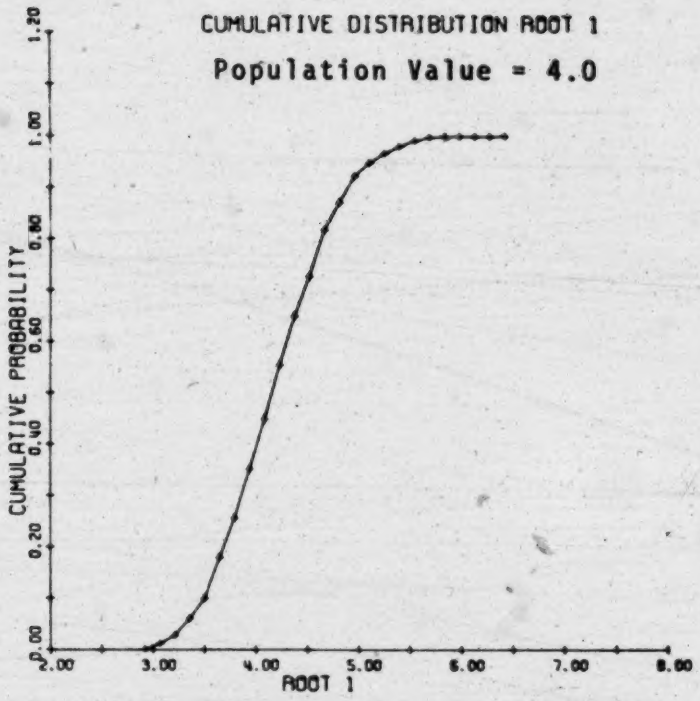
CUMULATIVE DISTRIBUTION ROOT 1
Population Value = 4.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
2.00
3.00
4.00
5.00
6.00
7.00
8.00
ROOT 1

Fig. 3.1

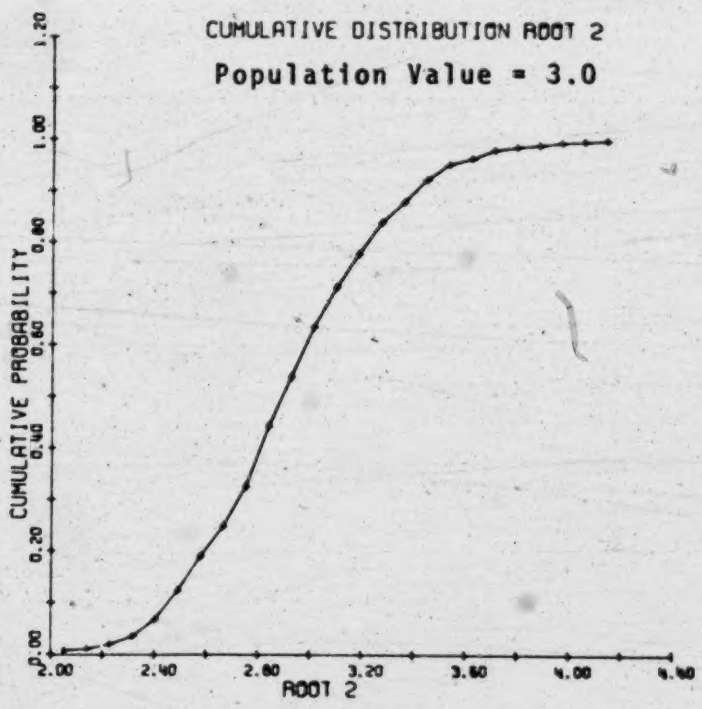
CUMULATIVE DISTRIBUTION ROOT 2
Population Value = 3.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
2.00
2.40
2.80
3.20
3.60
4.00
4.40
ROOT 2

Fig. 3.2

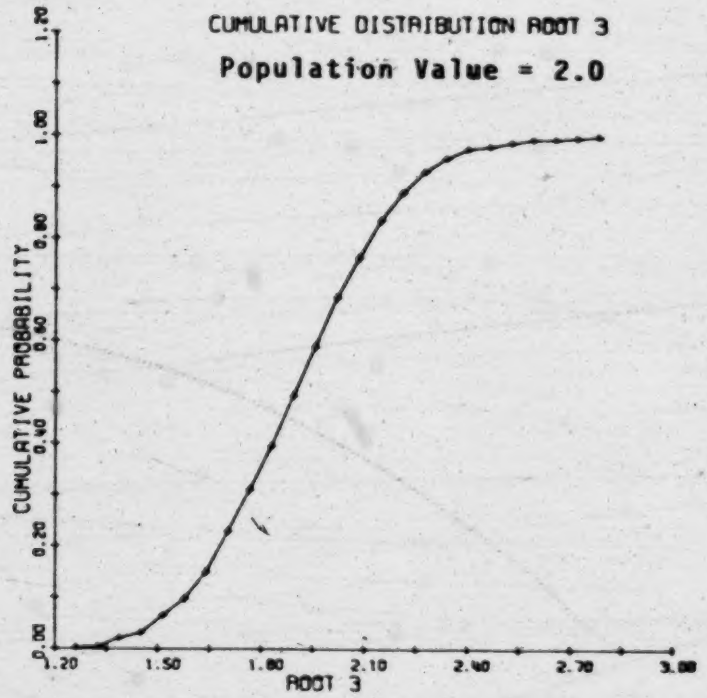
CUMULATIVE DISTRIBUTION ROOT 3
Population Value = 2.0
CUMULATIVE PROBABILITY
1.20
1.80
0.80
0.40
0.20
0.00
1.20
1.50
1.80
2.10
2.40
2.70
3.00
ROOT 3

Fig. 3.3

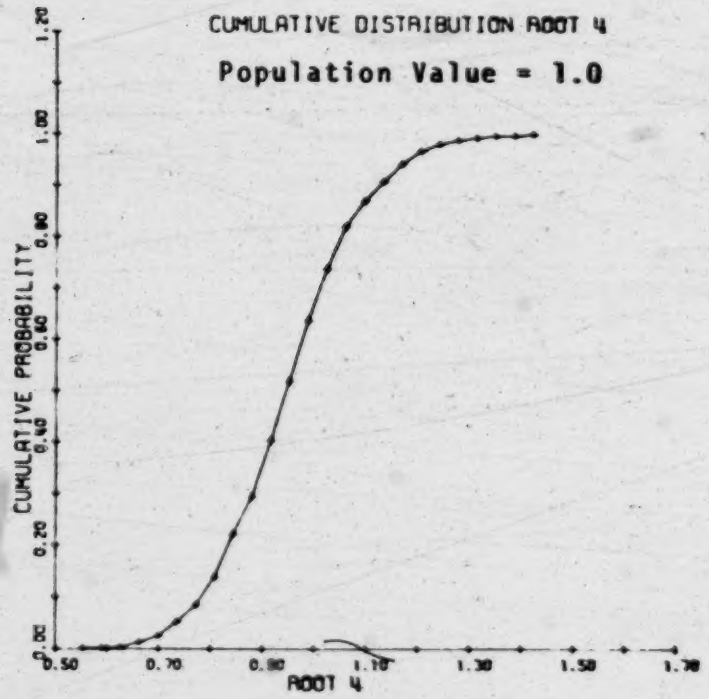
CUMULATIVE DISTRIBUTION ROOT 4
Population Value = 1.0
CUMULATIVE PROBABILITY
1.20
1.80
0.80
0.40
0.20
0.00
0.50
0.70
0.90
1.10
1.30
1.50
1.70
ROOT 4

Fig. 3.4

Computer Run 4

The Population Covariance Matrix

$$C = \begin{bmatrix}
0.938 & & & & & & & & & \\
0.748 & 1.064 & & & & & & & & \\
0.627 & 1.005 & 0.977 & & & & & & & \\
0.458 & 0.727 & 0.666 & 1.168 & & & & & & \\
0.649 & 0.918 & 0.810 & 0.308 & 1.359 & & & & & \\
0.227 & 0.383 & 0.294 & 0.590 & 0.536 & 0.614 & & & & \\
0.331 & 0.521 & 0.419 & 0.544 & 0.784 & 0.649 & 0.740 & & & \\
0.878 & 1.043 & 0.935 & 0.896 & 0.885 & 0.528 & 0.636 & 1.151 & & \\
0.080 & 0.697 & 0.751 & 0.489 & 0.526 & 0.197 & 0.275 & 0.511 & 0.879 & \\
0.931 & 0.994 & 0.883 & 0.828 & 0.802 & 0.444 & 0.544 & 1.116 & 0.405 & 1.110
\end{bmatrix}$$

TABLE 5

MEAN, VARIANCE AND SKEWNESS OF THE

NON-ZERO ROOTS OF RUN 4

Pop.	Root	Sample Mean	Sample Variance	$E(\ell_r)$	$Var(\ell_r)$	Skewness
7		7.0351	1.0011	7.0353	0.9898	0.1549
1		1.1949	0.0168	--	--	0.6369
1		0.9732	0.0089	--	--	0.1792
1		0.7820	0.0091	--	--	0.1060

Figs. 4.1, 4.2, 4.3 amd 4.4 show the cumulative distribution of these roots.

Fig. 4.5 show the cumulative distribution of roots 2, 3 and 4.

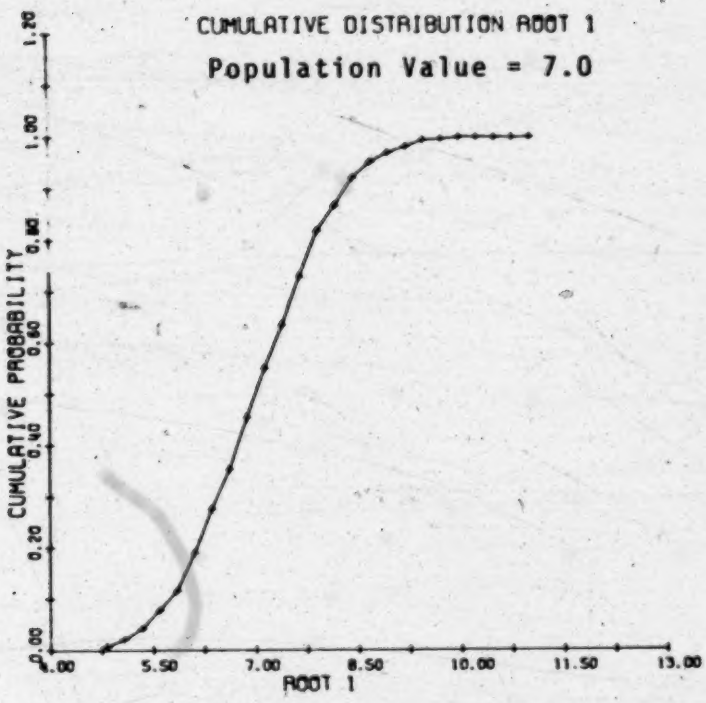
CUMULATIVE DISTRIBUTION ROOT 1
Population Value = 7.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
4.00
5.50
7.00
8.50
10.00
11.50
13.00
ROOT 1

Fig. 4.1

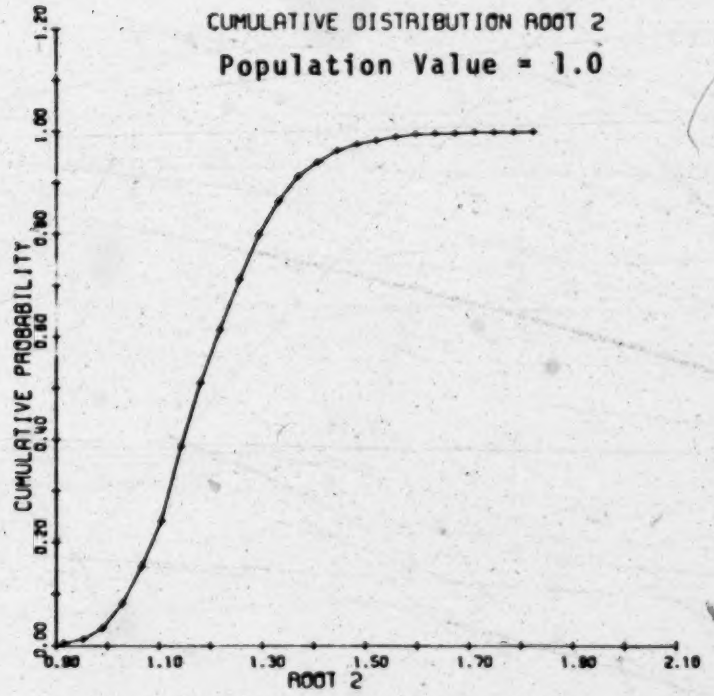
CUMULATIVE DISTRIBUTION ROOT 2
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.90
1.10
1.30
1.50
1.70
1.90
2.10
ROOT 2

Fig. 4.2

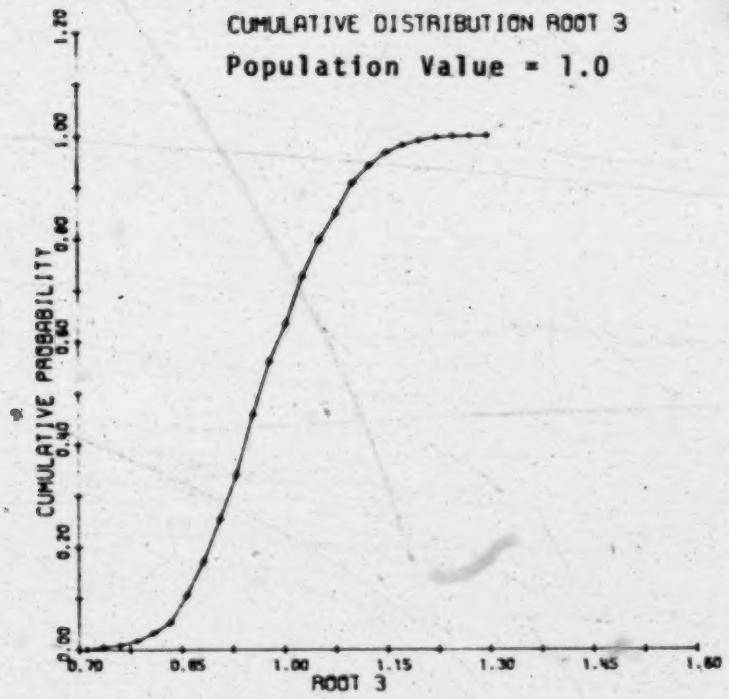
CUMULATIVE DISTRIBUTION ROOT 3
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.70
0.85
1.00
1.15
1.30
1.45
1.60
ROOT 3

Fig. 4.3

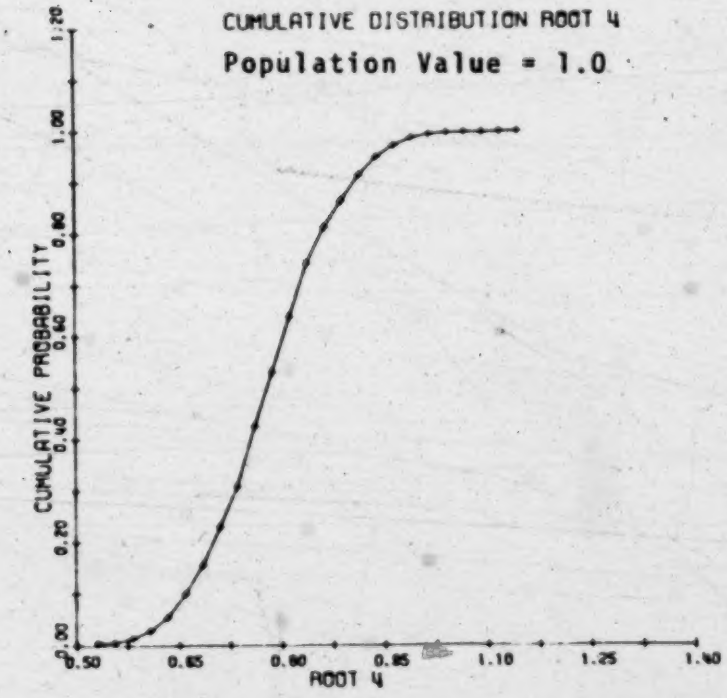
CUMULATIVE DISTRIBUTION ROOT 4
Population Value = 1.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.50
0.65
0.80
0.95
1.10
1.25
1.40
ROOT 4

Fig. 4.4

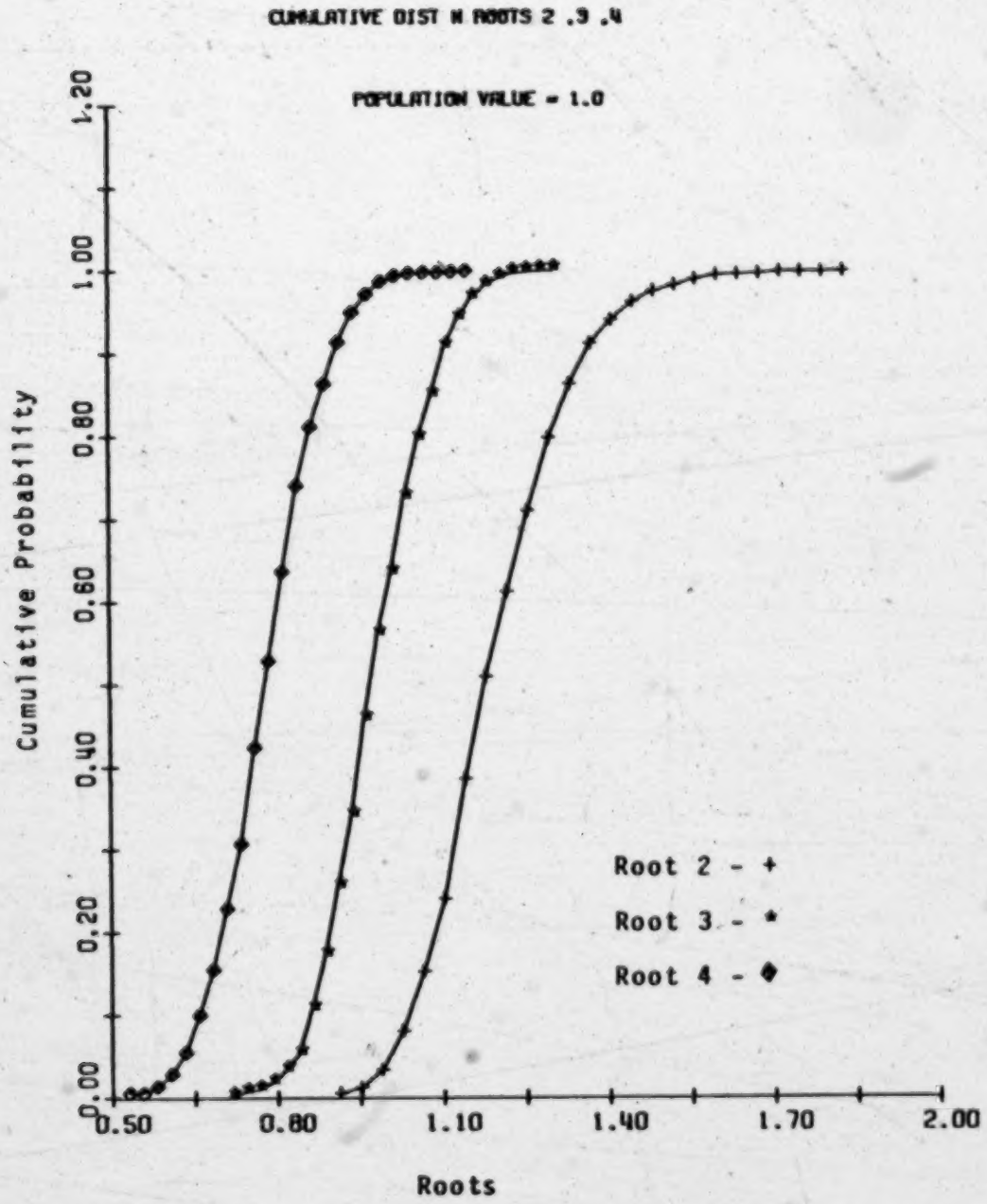
CUMULATIVE DIST N ROOTS 2 ,3 ,4
POPULATION VALUE = 1.0
Cumulative Probability
0.00
0.20
0.40
0.60
0.80
1.00
1.20
0.50
0.80
1.10
1.40
1.70
2.00
Roots
Root 2 - +
Root 3 - *
Root 4 - ◆

Fig. 4.5

Computer Run 5

The Population Covariance Matrix

$$C = \begin{bmatrix}
1.084 & & & & & & & & & \\
0.429 & 0.702 & & & & & & & & \\
0.287 & 0.712 & 0.773 & & & & & & & \\
0.064 & 0.312 & 0.292 & 1.421 & & & & & & \\
0.316 & 0.519 & 0.422 & -0.439 & 1.501 & & & & & \\
-0.099 & 0.026 & -0.087 & 0.587 & 0.387 & 0.842 & & & & \\
-0.020 & 0.127 & 0.003 & 0.355 & 0.721 & 0.821 & 0.892 & & & \\
0.633 & 0.580 & 0.500 & 0.584 & 0.379 & 0.272 & 0.308 & 0.710 & & \\
-0.455 & 0.569 & 0.751 & 0.317 & 0.291 & -0.034 & 0.194 & 0.150 & 1.280 & \\
0.799 & 0.562 & 0.468 & 0.518 & 0.287 & 0.146 & 0.172 & 0.726 & -0.015 & 0.794
\end{bmatrix}$$

TABLE 6

MEAN, VARIANCE AND SKEWNESS OF THE

NON-ZERO ROOTS OF RUN 5

Pop.	Root	Sample Mean	Sample Variance	$E(\ell_r)$	$Var(\ell_r)$	Skewness
4		4.1244	0.3144	4.1212	0.3136	0.2054
2		2.3652	0.0634	--	--	0.5623
2		1.9299	0.0351	--	--	0.1781
2		1.5521	0.0355	--	--	0.1300

Figs. 5.1, 5.2, 5.3 and 5.4 show the cumulative distribution of these roots.

Fig. 5.5 show the cumulative distribution of roots 2, 3 and 4.

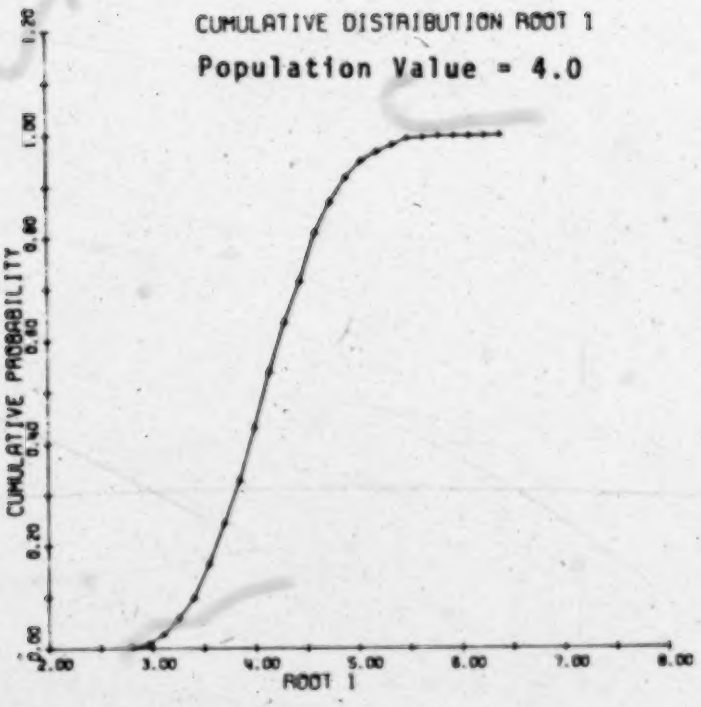
CUMULATIVE DISTRIBUTION ROOT 1
Population Value = 4.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
2.00
3.00
4.00
5.00
6.00
7.00
8.00
ROOT 1

Fig. 5.1

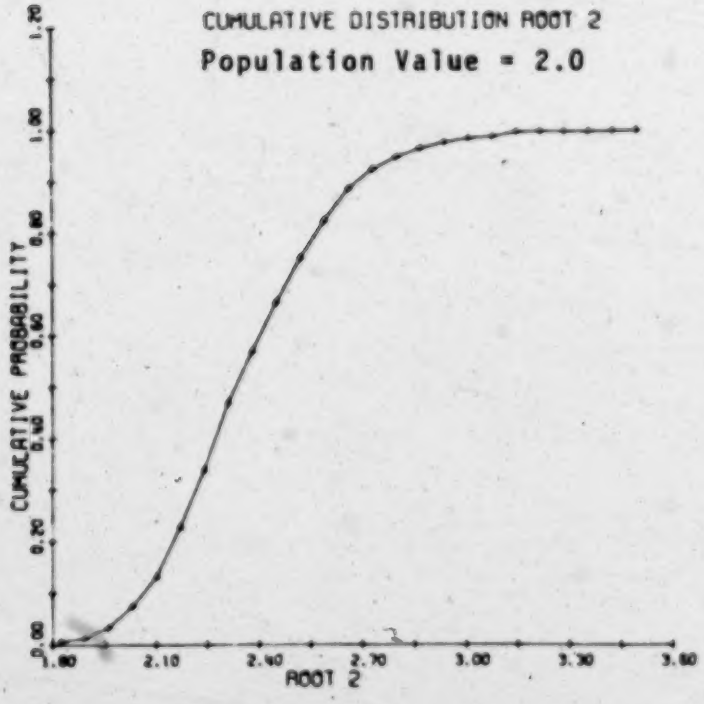
CUMULATIVE DISTRIBUTION ROOT 2
Population Value = 2.0
CUMULATIVE PROBABILITY
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.80
2.10
2.40
2.70
3.00
3.30
3.60
ROOT 2

Fig. 5.2

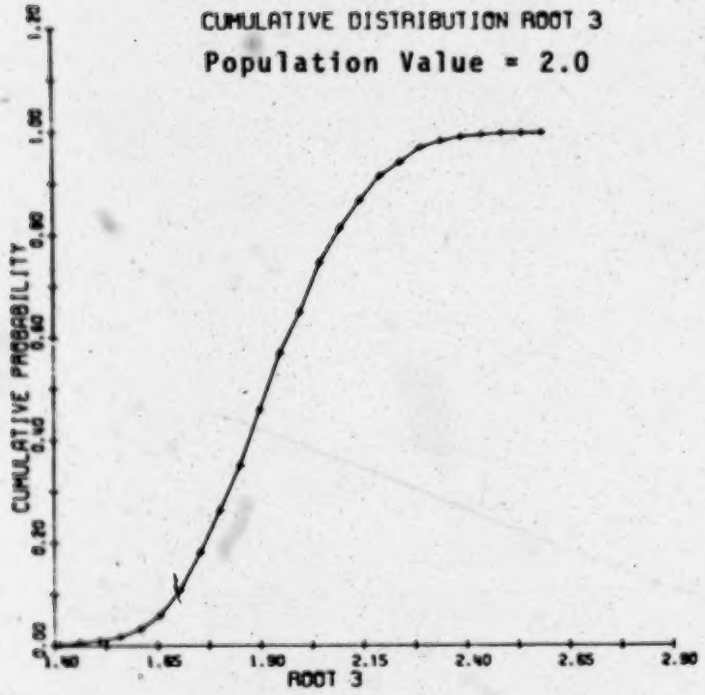
CUMULATIVE DISTRIBUTION ROOT 3
Population Value = 2.0
CUMULATIVE PROBABILITY
1.60
1.85
1.90
2.15
2.40
2.65
2.90
ROOT 3

Fig. 5.3

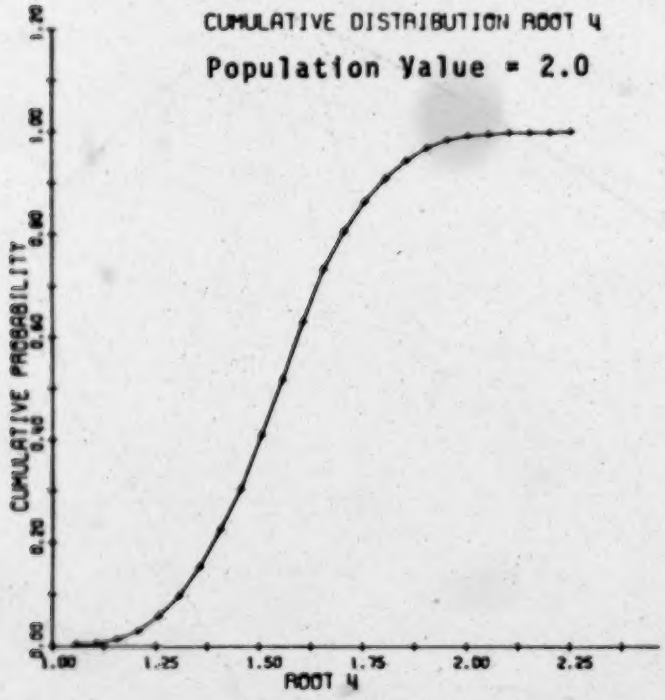
CUMULATIVE DISTRIBUTION ROOT 4
Population Value = 2.0
CUMULATIVE PROBABILITY
1.00
1.25
1.50
1.75
2.00
2.25
ROOT 4

Fig. 5.4

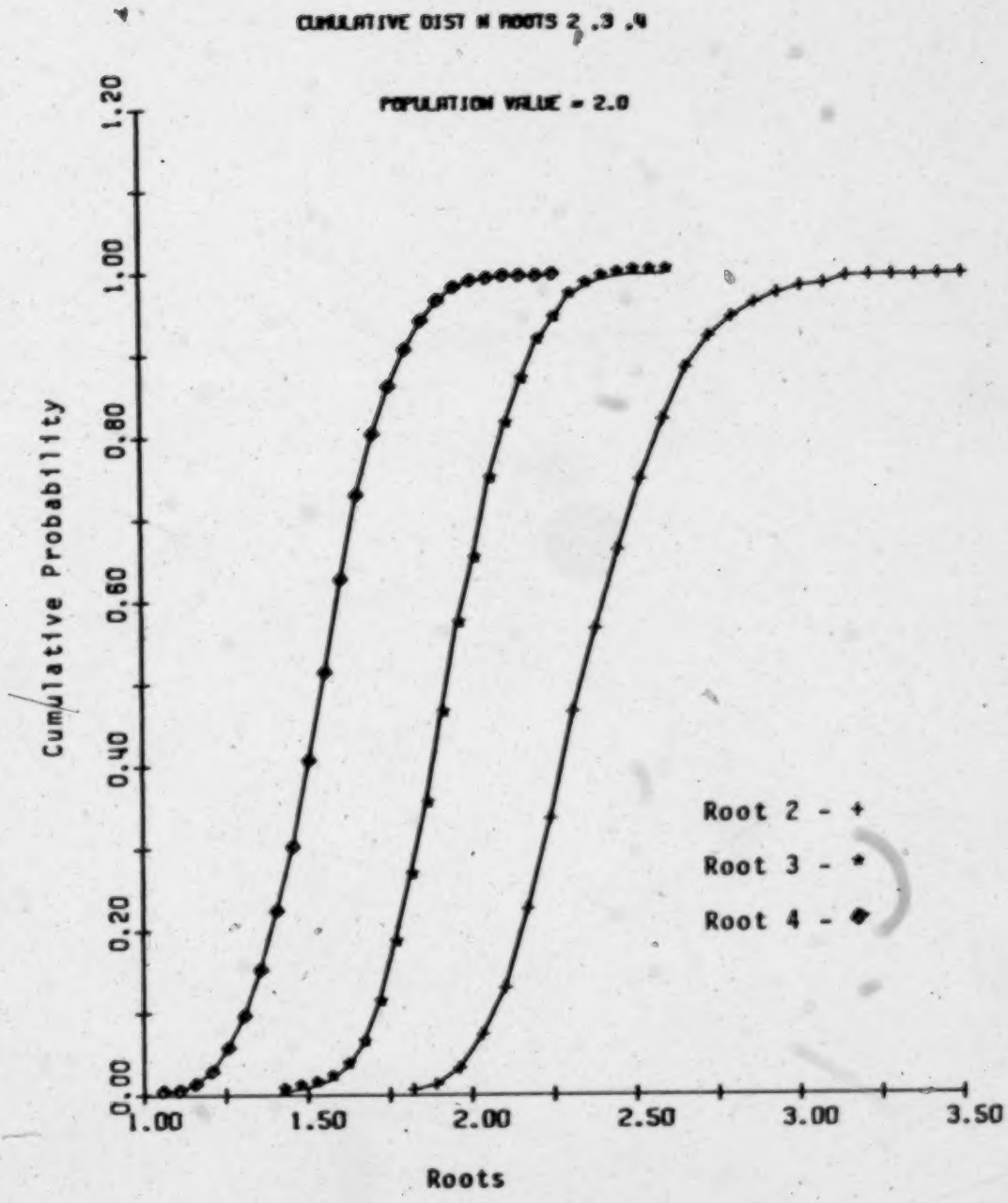
CUMULATIVE DIST N ROOTS 2 ,3 ,4
POPULATION VALUE = 2.0
Cumulative Probability
0.00
0.20
0.40
0.60
0.80
1.00
1.20
1.00
1.50
2.00
2.50
3.00
3.50
Roots
Root 2 - +
Root 3 - *
Root 4 - ◆

Fig. 5.5

CHAPTER 4

CONCLUSIONS

4.1 Discussion of Results

The values of $E(\ell_r)$ and $Var(\ell_r)$ for the k distinct non-zero roots using the expression given by Lawley were consistent with the values obtained using the sample mean and sample variance. The maximum discrepancy occurred for root 2 and run 3, where the sample mean calculated from the sample is 2.9189 and that calculated from Lawley's expression is 2.9545 giving a difference of 0.0356 from the sample or a discrepancy of 1.2 percent. The variance in this case is 0.1336 from the sample and 0.1446 from Lawley's expression. The difference here is 0.011 or 7.9 percent greater than the sample value.

Lawley's expression is applicable for the k distinct roots, but as seen from root 4 of run 1, all of the non-zero roots need not be distinct, the expression is also applicable where the non-zero root considered is distinct from the other roots.

The equal roots show substantial variability in mean and skewness. This variability may be a function both of the size of the root and the number of equal roots in the population.

From run 4 where the three population roots are equal

and the population value = 1.0, the sample roots range from 1.19 to 0.78 and in run 5 where the population value is 2.0, the values of the sample roots range from 2.37 to 1.55. For the case of two equal roots, when the population root = 2.0, the sample roots are 2.22 and 1.73 and for a population root of 1.0, the sample roots are 1.09 and 0.85 respectively. (The mean value of the sample root is used in each case.)

Figs. 4.5, 5.5, 1.5 and 2.5 show the variability in the cumulative distribution of these roots.

This variability is also very marked in the skewness of the distribution of the equal roots. The first of the equal roots is very positively skewed, and there is a sharp drop from the first to the second. This is noticeable in all cases whether there are two equal roots or three equal roots. In the case of three equal roots, the difference in the skewness of the second and third roots is less substantial. The skewness decreases with the rank order of the root, this is also the case with the sample means of the equal roots.

Since the number of equal roots is small, no generalizations can be made concerning the variability of the sample roots, no obvious relationship can be seen between the sample roots and the parent population root, however, there seems to be some relationship among the roots themselves.

In run 4 where the value of the population root is 1.0, the value of the sample roots are 1.10, 0.97, and 0.79. The second root is 81.5 percent of the first and the third is 81.4 percent of the second. In run 5 where the value of the

population root is 2.0, the sample roots are 2.365, 1.93 and 1.55 respectively. The second root is 81.6 percent of the first and the third root is 80.3 percent of the second, so that the relationship among the roots for the case of three equal population roots is fairly consistent.

In runs 1 and 2 where there are two equal roots in the population matrix, the second sample root is 78 percent of the first in both cases.

The relationship seems to be independent of the size of the population root but is dependent on the number of equal roots. The value of the sample root which corresponds to one of the equal roots of a population covariance matrix is therefore a function of the value of the population root, the rank order of the root among the equal roots, and the number of equal roots.

In order to establish some concrete results concerning the relationship among the sample roots when the population roots are equal, it would be necessary to obtain many samples for different numbers of equal roots in the population matrix.

In most tests where the sample covariance matrix is used, the population covariance matrix is usually unknown. Since the equality of the roots in the population is not evident in the sample roots, tests which are functions of a single root of the sample may have somewhat misleading results, unless a test is first made to determine the presence of non-zero equal roots in the population covariance matrix.

4.2 Comparison of Some Distributions

The Kolmogorov-Smirnov two sample test was used to test whether sample roots corresponding to population roots of the same numerical value and rank order, but where the remaining roots varied, were drawn from the same population or from populations with the same distribution.

TABLE 7

COMPARISON OF SOME DISTRIBUTIONS

Pop. Value	Run	Run	Rank Order of Root	Max. Dev.
5.0	1	2	1	0.019
4.0	3	5	1	0.049
3.0	2	3	2	0.035
1.0	1	3	4	0.018
1.0	1	2	4	0.330
2.0	1	5	2	0.252
2.0	1	3	3	0.295
1.0	2	4	4	0.271
1.0	2	3	4	0.346

The critical value of the Kolmogorov-Smirnov test at the five percent level of significance is 0.061. In the first four cases the maximum deviation is below this value and the hypothesis that these roots are from populations with the same distribution is accepted at the five percent level of significance.

The deviation in each of the remaining five cases is very significant, and the hypothesis of equal populations is therefore rejected at the five percent level of significance.

The results show that the roots which are distinct from all other non-zero roots in the population have the same distributions, while the non-distinct roots are from populations with significantly different distributions.

4.3 A Comparison of the Mean, Variance and Skewness of the Same Rank Order and Numberical Value for Four Different Runs

TABLE 8

MEAN, VARIANCE AND SKEWNESS OF THE 4th ROOT POPULATION VALUE = 1.0

Run	Pop. Roots	Mean	Variance	Skewness
1	5,2,2,1	0.947	0.018	0.171
2	5,3,1,1	0.848	0.012	0.144
3	4,3,2,1	0.951	0.018	0.205
4	7,1,1,1	0.782	0.009	0.106

When the roots are distinct from the other non-zero roots, that is in runs 1 and 3, the mean and variance are close enough to be considered equal, the only discrepancy in this case is the skewness of the distribution. The distri-

bution of the root from the population with all the non-zero roots distinct (run 3) is more positively skewed than the other. The other results from runs 2 and 4 where the root considered is one of the equal roots, the mean and variance in the case where there are two equal roots are greater, and the distribution more positively skewed than in the case where the root is one of three equal roots.

The results given in this thesis show the marginal distributions of the individual roots of a matrix with a Wishart distribution in the central case. The density functions of these roots are still unknown, both in the central and non-central cases.

Since these roots are used in various tests of significance, it is necessary that the sampling distributions should be known so that one would be aware of possible sampling errors and the power of the tests could be determined.

BIBLIOGRAPHY

1. Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York.

2. Anderson, T.W. (1963). A Symptotic Theory for Principal Component Analysis. Annals of Mathematical Statistics, Vol. 34, pp. 123-148.

3. Bartlett, M.S. (1950). Tests of Significance in Factor Analysis. British Journal of Psychology, Statistical Section, Vol. 3, pp. 77-85.

4. Birnbaum, Z.W. (1952). Numerical Tabulation of the Distribution of Kolmogorov's Statistic for Finite Sample Size. Journal of the American Statistical Association, Vol. 47, pp. 425-441.

5. Box, G.E.P. and Muller, M.E. (1958). A Note on the Generation of Random Normal Deviates. Annals of Mathematical Statistics, Vol. 29, pp. 610-611.

6. Chen, E.H. (1971). A Random Normal Number Generator for 32-Bit Word Computers. Journal of the American Statistical Association, Vol. 62, pp. 607-625.

7. Graybil, F.A. (1969). Introduction to Matrices with Application in Statistics. Wadworth Publishing Co., Inc., California.

8. Kullback, S. (1959). Information Theory and Statistics. John Wiley and Sons, Inc., New York.

9. Lawley, D.N. (1956). Tests of Significance for the Latent Roots of Covariance and Correlation Matrices. Biometrika, Vol. 43, pp. 128-136.

10. Lawley, D.N. and Maxwell, A.E. (1971). Factor Analysis as a Statistical Method, [2nd Ed.]. Butterworths, London.

11. Massey, F.J. Jr. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, Vol. 46, pp. 68-77.

12. Morrison, D.F. (1967). Multivariate Statistical Methods. McGraw-Hill Book Co., New York.

13. Naylor, T.H., Balintfy, J.L., Burdick, D.S. and Chu, K. (1966). Computer Simulation Techniques. John Wiley and Sons, Inc., New York.

14. Newman, T.G. and Odell, P.L. (1971). The Generation of Random Variates. Charles Griffin and Co. Ltd., London.

15. Ralston, A. and Wilf, S.H. (1967). Mathematical Methods for Digital Computers, Vol. II. John Wiley and Sons, Inc., New York.

16. Scheuer, E.M. and Stroller, D.S. (1962). On the Generation of Normal Random Vectors. Technometrics, Vol. 4, pp. 278-281.

17. Siegal, S. (1956). Non-Parametric Statistics for the Behavioral Sciences. McGraw-Hill Book Co., Inc., New York.